生物多样性语义知识抽取研究探索*

刘建华[1,2] 王颖[1] 张智雄[1] 李传席[1]

[1]（中国科学院文献情报中心 北京 100190）

[2]（中国科学院大学 北京 100190）

摘要:

[目的]拓展以物种为中心的生物多样性知识抽取框架，探索实现语义知识抽取方法

[方法]结合当前生物多样性抽取的主流研究，以物种为中心，设计包含多种实体及实体间关系的知识抽取框架，利用已有的众多专业数据库，设计并实现相应的识别方法。

[结果]设计了以物种为核心的知识抽取框架，探索实现了多种实体及实体间关系的语义知识抽取方法，拓展了生物多样性领域抽取内容和思路。

[局限]本研究实体识别的完整性和准确性受底层知识库影响较大，且实体间关系的类型局限于共现、上下位类、语法关系几类，还需进一步研究。

[结论] 拓展了生物多样性领域抽取内容和思路，可有效支持后续的语义检索、科学计算。

关键词：生物多样性 物种 知识抽取 关系识别

分类号：G250

Study on Semantic Knowledge Extraction in Biodiversity

Liu Jianhua[1,2] Wang Ying[1] Zhang Zhixiong[1] Li Chuanxi[1]

[1](National Science Library, Chinese Academy of Sciences, Beijing 100190, China)

[2](University of Chinese Academy of Sciences, Beijing 100190, China)

Abstract:

[Objective] It is aimed to expand the knowledge extraction framework in biodiversity, and implement the knowledge extraction method.

[Methods] This paper designs a knowledge extraction framework with various entities and entity relationships for biodiversity，which combines with current main biodiversity extraction research and takes species as the center. Besides, it implements the knowledge extraction method based on amount of specialized databases.

[Results] This paper designs a species-central knowledge extraction framework for biodiversity, implements the knowledge extraction method about semantic named entities identification and relationships identification among them, and expands the units and methods for knowledge extraction in biodiversity field.

[limitations] The recall and precision of the knowledge are effected by the dictionaries and rules. Besides, the semantic relationships among named entities are limited in co-occurrence, hierarchical and simple syntactic relationships.　All mentioned above should be improved in the future.

[Conclusions] It expands the knowledge units and methods for knowledge extraction in biodiversity. It could support the follow-up semantic retrieval and computation.

Keywords：biodiversity science; species; knowledge extraction; relation extraction;

---

* 本文系国家十二五科技支撑计划项目“面向外文科技文献信息的知识组织体系建设与应用示范(STKOS)”的子课题“信息资源自动处理、智能检索与 STKOS 应用服务集成”(项目编号:2011BAH10B05)成果之一

1 引言

随着全球气候变暖、各种自然灾害频发等问题，物种灭绝速度越来越快，针对生物多样性保护与持续利用的研究日益成为生物多样性研究的焦点，大量与之相关的研究论文急剧增长。如何帮助科研人员从这些富含了大量物种名称（科学命名、别名、俗名、变种名等）、基因、实验设备等实体的文档中快速发现所需信息，是生物多样性信息学面临的重要问题之一。针对此，越来越多的研究者正努力尝试利用现有众多的生物多样性专业数据库，如物种名录、标本库、图片库、基因库等，从生物多样性描述文本或文献中提取知识对象，并借助语义内容标注技术实现知识对象的自动深层标引，实现数字资源之间的语义集成和关联，从而为进一步的语义检索、数据挖掘、科学计算提供支撑。

本文在对当前生物多样性信息抽取领域相关研究分析的基础上，结合中国科学院文献情报中心"建设生物多样性领域本体构建与语义组织应用示范平台"的实际要求，设计了生物多样性语义知识抽取框架，探索实现了相应的语义知识抽取方法，开发了相应的生物多样性示范平台。

2 相关研究概述

在众多研究者的努力下，目前已经出现了不少针对生物多样性领域的信息抽取工具，这些工具或者采用单一的自然语言处理、词典、机器学习、规则模板、浅度或深度句法解析、概率分类等方法，或者融合上述几种方法进行识别，识别的内容多数集中于物种的各类名称（科学命名、别名、俗名、变种名等），部分工具涉及对物种的性状的识别。Anne E. Thessen 等人在文献 1 中[1]综述了当前在生物多样性领域使用自然语言处理和机器学习算法实现物种名称识别的的相关研究，Nona Naderi 等人在文献 2 中[2]集中介绍了 GATE 框架下提供的生物医学领域的各种工具。上述文献对常规的生物多样性信息抽取流程，主流的信息抽取方法（基于词典、基于规则、浅度句子解析、深度句法解析等）进行了全面的评述，并对各个阶段的主要信息抽取工具进行了全面的综述。本文将不再对上述内容进行重复介绍，而是结合当前一些重要的生物多样性信息抽取工具，重点对生物多样性领域抽取的内容进行探讨，希望在此基础上对笔者进一步提出生物多样性知识抽取框架提供参考支撑。

在目前生物多样性抽取研究中，主要抽取内容可以归纳为以下几个方面：

2.1 物种名称识别及规范

由于语种、地方称谓等的差异，科技文献中出现的同一个物种名称是多种多样的。有的是标准规范的双名制命名法（或三名制命名法）形成的拉丁文名，即属名加种名（若是亚种则在种名前再加上一个亚种名），且属名在前，种名在后，属名第一个字大写，种名小写，属种名称均为全称，后面通常还会跟随着物种命名人的姓氏[3]；有的采用取属名首字母、种名全称的缩写方式；有的会采用物种的俗名（可能是英文，也可能是其他语种，同一个物种在不同的国家或地区也可能会有不同的俗名）[4]。这些问题的存在大大增加了物种名称识别的难度。因此目前有不少研究者专门针对物种名称的识别、规范及组织进行了研究，这也是当前生物多样性抽取相关研究的主流。这些研究中产生的研究成果中比较典型的包括可用作物种名称识别与规范词典的 NCBI taxonomy[5]、BioNames[6]（一个将动物名称与其来源描述、分类及进化树关联的在线数据库）、物种 2000 全球生物

物种名录[7]，也包括各种比较成熟物种名称识别工具如 NetiNeti[8]、Linnaeus[4]、OrganismTagger[9]、TaxonGrab[10]等。

2.2 物种性状识别

对物种分类学研究人员而言，物种的各类性状描述信息，如根、茎、叶的颜色、长度等，是界定物种门类的重要参考信息。因此，有一部分生物信息学研究人员着力于探索物种各类性状的自动识别方法。Taylor[11]在分析文本语法特征的基础上, 以人工方式建立规则和词典,实现了物种部位、特征及状态等描述信息的识别。Tang[12]等在相关研究基础上，通过预定义模板的方式，有监督地学习生成相关的规则，实现了对物种叶子的形状、大小、颜色、排列及果实的形状特征的识别。Hong Cui[13]等人开发的 CharaParser 采用启发式方法和句法特征生成规则，较好地实现了对物种多类性状的识别。段宇锋等人[14]持续探索着中文植物物种多样性描述文本中形态信息的抽取。

2.3 生物网络识别[15]

各种生物实体（物种、分子、基因、蛋白等）之间存在着多种关系，这些关系可以用网络图的方式表达出来，进而通过对图的分析实现对生物系统的分析[16]。蛋白质和基因是生物医药领域普遍关注的重点内容，关于这类知识的识别研究并不限在生物多样性领域开展。当前生物多样性相关的文献中，研究人员可通过对物种基因测序的方式来鉴定物种的亲缘性，也可通过采用蛋白质或基因技术影响或改变生物的内外环境或特征，从而研究相关问题。因此，对蛋白质和基因的识别更多地不仅仅是识别出蛋白质、基因等命名实体单元，而是识别出各类生物实体之间通过动词（或动词短语）、介词（或介词短语）、所有格等关联而成的生物网络关系。在此基础上可进一步开展资源的重组织、语义检索、计算分析等工作。

3 语义知识框架设计

上文对当前生物多样性抽取领域当前重点关注的抽取内容及其相关的资源工具进行了分析。结合中国科学院文献情报中心“建设生物多样性领域本体构建与语义组织应用示范平台”的实际要求，从实际应用的角度出发，在人工标引了100 篇生物多样性领域的科技文献后，笔者以物种为核心，综合分析了当前生物多样性领域研究中可能涉及的与物种研究相关的知识单元类型，各知识单元类型之间的关联关系，设计了如图 1 所示的生物多样性语义知识框架，该语义知识框架是进一步支持笔者开展知识抽取、知识组织的基础。

从图中可以直接看出，笔者的知识框架中包含了两个方面的语义知识：语义知识单元、语义知识单元之间的语义关联。

3.1 语义知识框架中的知识单元

这里的知识单元即图 1 每个文本框中列出的语义类型，在实际的科技文献中，这些知识单元往往以命名实体名称或短语的形式表达出来，将科技文献中提及的命名实体名称或短语以图 1 中定义的语义类型进行标注，即可实现该语义单元的识别。图 1 中所有的知识单元语义类型均以图中心的物种为核心，这些类型覆盖了物种的各个方面，包括名称、分布、特征、生长发育阶段、影响因素等，部分大的分面上还有其进一步细分的下级类，部分分面会共有一些语义单元。

- 物种名称。包括各种物种名称、变种名称、品种名称、变型名称、物种的各种俗名。
- 物种特征。包括各类物种的器官、细胞、基因等。
- 物种分布。这里的物种分布包括地理区域上的分布，同时还包括不同生态环境下的分布，因此，该方面的知识单元除了洲、国、地区、城市、县等地理名称外，还包含生物群落、地貌、物理环境（高度、温度、湿度等）。
- 物种生物发育阶段。包括物种的发育阶段、物种各器官的发育阶段。
- 对物种产生影响的因素。能对物种产生影响的包括非生物因素和生物因素两类，其中，非生物因素包括温度、湿度、海拔高度、土壤等，生物因素则包括各种细胞、染色体、蛋白质、DNA、基因片段、化学元素、化合物等。
- 对物种分类的各种标准和生态位模型工具。
- 对物种实验的各种分析方法及设备仪器。
- 其它基本信息。包括人、机构及目前无法确定明确语义的名词短语。

这些知识单元基本上涵盖了当前生物多样性，尤其是物种多样性研究中的主体知识单元，它们构成了相关研究的主要知识点。

3.2 语义知识框架中的语义关联

上文中分析的这些知识单元并不是以独立的形式存在于科技文献中，他们彼此之间往往还存在着各种语义关联，结合这些语义关联才能够最大化地利用这些知识单元实现深层的文本内容挖掘。在本文定义的语义知识单元中，笔者根据实际应用及后续能够识别出来的现实情况，定义了有限的几种语义关联，这些语义关联可以作为事实三元组支持进一步的文本分析。例如：

<生物因素/非生物因素>在<物理环境>下

<生物因素/非生物因素>作用于<物种/器官/细胞>

<生物因素/非生物因素>作用于<生物阶段>

<分析方法/仪器设备>作用于<物种/器官/细胞>

<物种/器官/细胞>呈现的<生物特征>

<物种/器官/细胞>的<生物阶段>

<物种>分布于<分布区域>

<分布区域>的<地貌、植被、土壤等特征>

3.3 其它

除了上述在知识框架图中明显展示出来的两个方面的语义知识外，笔者注意到，在实际的科技文献中还存在不少有分析价值的语义标注。根据人工标引的科技文献，笔者发现，有不少知识无法简单地以某个知识单元或某个知识单元间关联关系的形式展示出来，比如一个完整的实验条件（如化学元素的浓度与温度控制综合作用的实验条件）、一个完整的实验过程等，这些知识可能包含了多个知识单元和知识单元间的关联关系。针对这些内容，笔者可以采用知识句群的方式进行表达，即将关联密切的多个短语或短句组织在一起，以保证知识的完整性。依据他们的内容，可以简单将这类知识划分为：方法、过程、结果几类。这些内容与上述的两类语义知识共同构成了生物多样性语义知识框架。针对这一部分知

识的识别方法将在后续的研究中进一步阐述，下文将围绕前两类知识的识别展开实验探索。

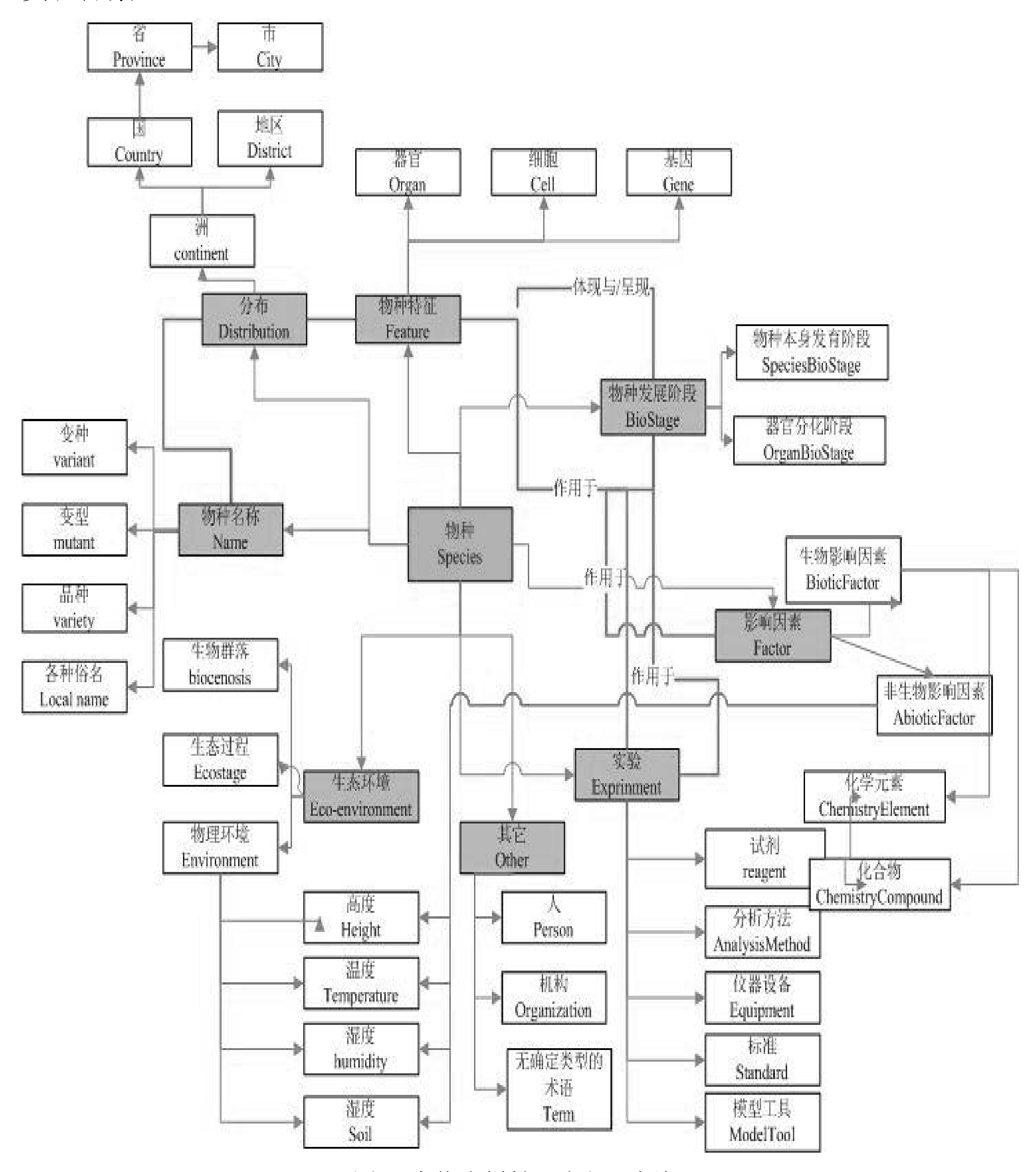


图 1 生物多样性语义知识框架

4 语义知识抽取的实现

基于上文定义的生物多样性语义知识框架，笔者尝试利用词典、规则、句法分析等综合方法，从检索获取的生物多样性相关的科技文献摘要中，识别出知识框架中定义的知识单元和知识单元间的关联关系。

4.1 实验数据及语料的选择

为了探索生物多样性领域的知识抽取，笔者从 pubmed 数据库的 Plant Physiology、The Plant Cell 两个期刊上获取了 23000 篇左右的期刊文摘，并根据中国科学院植物研究所提供的 20 种核心期刊列表，从 WOS 获取了 27049 条科

技文摘数据。本研究将设计相应的方法来识别出这些摘要中提及的语义知识。为了提升本研究识别的效率，笔者通过专家咨询及参考中科院植物研究所的相关研究[17]，收集整理了可作为信息抽取词表的相关语料，主要包括：植物所提供的G2000 植物本体数据库、NCBI 物种库、UMLS 中的相关领域术语和词汇、地址名称词表、Chemical Entities of Biological Interest 中的小化合物名称等，这些领域资源将作为实体名称识别的重要支撑。

4.2 知识抽取框架的设计

为了更好地实现知识单元及知识单元间关系的识别，笔者设计了图 2 所示的知识抽取框架，具体步骤描述如下：

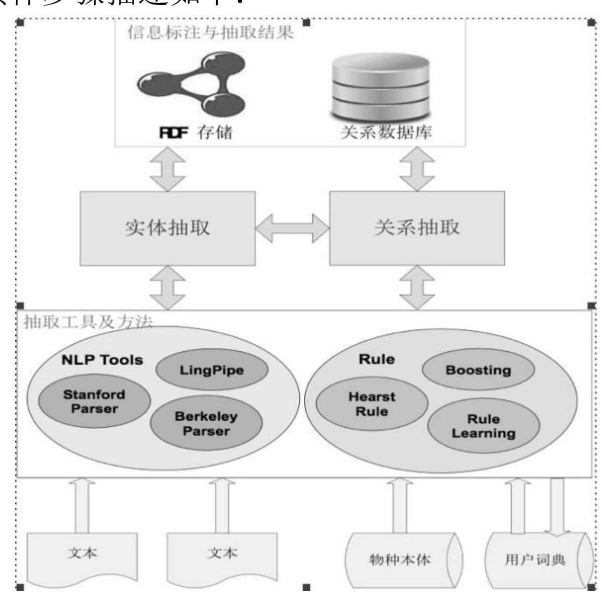


图 2 语义知识抽取框架

（1）输入数据源

主要包括待抽取的科技文献及相关领域资源（植物多样性本体、NCBI 物种库等）。

（2）抽取工具及方法

通过采用不同的自然处理工具（包括 Stanford Parser、Berkerly Parser 等），实现对文本的词性标注、句法依存关系分析及句子的语法语义分析。通过结合不同的抽取规则和距离度量算法，实现句子中的实体与关系的识别。

（3）实体抽取与关系抽取

实体抽取与关系抽取之间是一个交叉迭代实现的过程，一方面，实体抽取过程的本身是一个迭代过程，新识别的命名实体添加到用户词典中，用于下一轮的实体识别过程；另一方面，关系抽取的结果也可以用于发现新的实体，新发现的实体用于下一轮的关系发现过程。

（4）信息抽取结果存储

根据信息抽取结果类型的不同，采用 RDF 存储和数据库存储两种方式实现实体及关系的存储。

4.3 知识抽取的流程

（1）知识单元的标注与抽取

命名实体的识别主要方法包括基于词典和规则的方法，以及基于统计的方法等。在这里，笔者采用的命名实体识别方法以词典为基础，采用基于规则和统计方法相结合，实现新实体发现识别。

① **基于领域词典的实体标注。**对领域资源进行分析提取，形成可用于命名实体抽取的领域词典，实现对科技文献中所涉及的实体标注。在具体实现过程中，严格按照词典进行标注，获取实体在句子中的相关信息，如图 3 所示，彩色部分为标注结果。

② **基于词典相似性的新实体识别。**基于词典的命名实体识别无法解决未登录词的问题，通过识别文本中含有的命名实体，并计算其与词典中命名实体之间的距离，实现对一些未登录词的识别。对于上例中，Solanum section Petota 作为一个整体出现表示一个命名实体，而基于词典的方法则只识别了 Solanum，则可以通过相似性扩展，实现规范的实体识别，从而实现实体 sect. Petota 的识别。

Species boundaries were assessed by phenetic analyses of morphological data for all species of wild potatoes (Solanum section Petota) assigned to ser. Longipedicellata: S. fendleri, S. hjertingii, S. matehualae, S. papita, S. polytrichon, and S. stoloniferum. These six tetraploid species grow in the southeastern United States (S. fendleri) and Mexico (all six species). We also analyzed morphologically similar species in ser. Demissa (S. demissum) and ser. Tuberosa (S. avilesii, S. gourlayi, S, verrucosum). We chose S. verrucosum and S. demissum as Mexican representatives, and S. avilesii and S. gourlayi as South American representatives of other series that are difficult to distinguish from ser. Longipedicellata. We also analyzed morphologically more dissimilar species in ser. Tuberosa (S. berthaultii) and ser. Yungasensia (S. chacoense). The results support only three species in ser. Longipedicellata: (1) S. polytrichon, (2) S. hjertingii + S. matehualae, (3) S. fendleri + S. papita + S. stoloniferum. Solanum avilesii, S. gourlayi, and to a lesser extent S. demissum and S. verrucosum are very similar to members of ser. Longipedicellata and are difficult to distinguish practically from them, despite differences in chromosome numbers and crossability relationships. These data help document and explain the extensive taxonomic difficulty in sect. Petota, highlight conflicts between biological and morphological species concepts, and add to a growing body of evidence that too many wild potato species are recognized. (1192769)

图 3 基于领域词典的实体标注样例

③ **基于语法关系的新实体识别。**文本中出现的有些实体通过上述两种方法仍然无法辨识，例如 ser. Longipedicellata, ser. Tuberosa, ser. Yungasensia , S. matehualae。对于这些词的识别，通过分析句子的句法依存关系及语法关系（并列的句子成分），结合统计分析算法，可以实现命名实体的识别。

④ **实现文献中术语的标识。**除了文本中包含的命名实体，领域的术语具有提示文献内容的重要作用，因此术语的识别有助于为用户提供文献内容的直接简洁的认知。通过词法分析方法（名词词组等）对文献中出现的重要术语进行标注，如：Species boundaries, phenetic analyses, morphological data, tetraploid species 等。

⑤ **地理位置的识别。**对于此类可穷尽的地理位置信息，通过地理词典实现包括城市、国家等信息的识别，例如：Mexican , South American , United States。

⑥ **数字信息的识别。**主要是识别文本中含有的数值相关信息，如年份、日期、实验数据，以及相关的描述数值等，此类信息主要可借助构词法规则、特殊数值词典等实现。如对图 1 中的文本，可以识别出 six tetraploid species，three species。而对于文本“The inhibition constant values were 0.46 (using acetolactate as substrate) and 0.19 [mu]M (acetohydroxybutyrate), respectively. ”则可以识别其中所包含的 0.46 和 0.19 [mu]M。

⑦ **实体属性标注。**除了标注命名实体以外，对于识别出的命名实体的描述信息进行标注，可以更加全面的提示命名实体所包含的信息。通过分析命名实体出现的上下位语境信息（特定语法规则、句法依存规则等），可实现实体属性的标注。例如文献中含有词组 wild potatoes，命名实体识别可识别出 potatoes，通过 NP 名词组块的句法依存关系，可将 wild 标注为该实体的属性，从而为用户提供更为精确的信息。

（2）关系的抽取

① **多层级的共现关系。**在不同位置的共现关系可以用于计算实体之间的关联关系。本研究中重点考虑了命名实体在标题、摘要及句子级的共现关系，通过分析标注出的命名实体出现的位置，可以获取实体之间的共现关系。

- 句子级共现：

< S. fendleri, S. hjertingii >, < S. hjertingii , S. matehualae >,< S. verrucosum , S. demissum >,< Solanum avilesii, S. stoloniferum >,< S. fendleri , S. papita >…

- 摘要级共现：

< potatoes, S. berthaultii >, < S. stoloniferum , S. fendleri >,< S. fendleri , S. verrucosum >,< S. gourlayi , S. avilesii >,< S. demissum , S. fendleri >…

② **实体的同位语法关系抽取。**针对如上例子，得到如下结果：

同位语关系： < S. verrucosum , S. avilesii >, < S. gourlayi, S. avilesii >, < S. fendleri, S. matehualae >, <Solanum section Petota , wild potatoes >等。

③ **实体的并列语法关系抽取。**针对如上例子，得到如下结果：

并列关系：<S. matehualae , S. stoloniferum >,<S. polytrichon , S. hjertingii >,<S. fendleri , S. stoloniferum >,<S. hjertingii , S. papita >,<lesser extent S. demissum , S. verrucosum >,<S. hjertingii , S. stoloniferum >等。

④ **事实关系识别。**在标题、摘要中存在的<S, P, O>（主语，谓词，宾语）事实，可为后续的关系推理提供重要的支持，这一类的事实包括通用型事实与植物本体中定义的事实关系，借助于句法依存关系分析、本体映射，对上例（1192769）文本进行抽取，可以得到如下结果：

- <"We","also analyzed morphologically","similar species">
- <"These data","explain the extensive taxonomic difficulty in" ,"sect. Petota ">
- <"South American representatives of other series","are difficult to distinguish from"," Longipedicellata ">
- <"These six tetraploid species","grow in","the southeastern United States (S. fendleri) and Mexico">
- <"We","also analyzed morphologically","more dissimilar species" >

- <"Species boundaries","were assessed by","phenetic analyses of morphological data" >
- <"a lesser extent S. demissum and S. verrucosum","are very similar to","members of ser. Longipedicellata ">
- <"The results","support only","three species">

⑤ **语义上下位关系的发现。**通过采用基于规则的方法，可以发现术语之间的语义上下位关系，如下例所示：

- “CSS grass margins could be improved as butterfly habitats if they are linked to existing habitats such as hedgerows, are sown with a better range of native grasses and herbs and are managed in a way more conducive to wildlife.(1196577)”

可以标识出 hedgerows 是于 hedgerows 的下位术语，即<hedgerows, hypogyny , habitats >

- “We investigated all sections of genus Cochlearia recognised in the most common concepts, as well as some genera such as Ionopsidium, Bivonaea, Pastorea and Thlaspis.(1205921)”

可以标识出 Ionopsidium, Bivonaea, Pastorea 和 Thlaspis 是 genera 的下位术语，即<Ionopsidium, hypogyny , genera>，< Bivonaea, hypogyny , genera >，< Pastorea, hypogyny , genera >，<Thlaspis, hypogyny , genera>。

物种性状关系识别。例如：wild potatoes，可以标注出 potatoes 具有属性 wild，即<potatoes,have property , wild>。

4.4 知识抽取的结果应用

以领域词典和人工为主撰写的规则库为重要的知识库支撑的知识抽取方法虽然在领域快速迁移与新物种或新知识单元识别的灵活性方面有所欠缺，但是其准确性可以得到有效的保障，从而进一步支撑实际的知识检索应用。利用上述定义的知识抽取框架和抽取方法，笔者共计从 6 万多篇相关的文献标题和摘要中获得了 273,668 条知识单元的抽取结果，各抽取类型的分布结果如下。

基于上述知识抽取的结果，综合利用领域知识库和其它第三方资源，笔者进一步构建了生物多样性领域语义检索的应用示范平台，为用户提供领域知识揭示、语义标注、本体导航等检索应用。

对文献标题和摘要进行标注，从 64,475 篇文献中获得 273,668 标注条结果

实体类型	数量	实体类型	数量
genus	115698		
family	25332		
habit	13510		
plantFlowerColor	12649		
cultivatedHabitat	12277		
plantStemType	10306		
species	9478		
longevity	8233		
plantFruitType	6489		

plantGynoeciumCarpelFusion	4875		
planAndroeciumStamenArrangement	4793		
plantLeafArrangement	3908		
plantLeafShape	3609		
plantLeafMargin	3268		
plantInflorescenceForm	3268		
plantLeafSurface	2859		
plantNumbersOfFloralStructure	2815		
plantLeafDivision	2615		
lossOfLeaves	2482		
photosynthesis	2282		
plantFlowerSexuality	2222		
plantAndroeciumStamenType	2152		
plantStemForm	1983		
province	1845		
plantFlowerTime	1773		
plantRootType	1725		
plantInflorescenceType	1637		
plantPollinationSystem	1509		
gene	1270		
plantFlowerPerianthType	1227		
order	1088		
plantFlowerSymmetry	1043		
plantLeafApex	780		
plantAndroeciumAntherAttachment	736		
plantGynoeciumOvaryPosition	722		
plantAndroeciumStamenFusion	717		
plantFlowerPerianthForm	621		
plantLeafStructure	510		
plantLeafAttachment	323		
culturedHabitat	264		
phylum	252		
plantLeaf	244		
class	153		
plantRootStructure	127		
plantRootForm	77		

plantLeafDivision|plantLeafShape	73		
plantInsertionOfFloraStructure	73		
plantGynoeciumStyleForm	64		
plantGynoeciumCarpelNumber	62		
plantLeafVenation	56		
plantStemStructure	52		
plantGynoeciumCarpelType	48		
aquaticHabitat	39		
plantFruitStructure	35		
plantAndroeciumStamenNumber	22		
plantGynoeciumPlacentation	19		
plantFlowerStructure	17		
plantGynoeciumOvuleType	13		
plantInflorescenceStructure	9		
plantGynoeciumStructure	1		
plantFlowerPerianthStructure	1		
extremeHabitat	1		
plantFruitColor	1		

实体类型和数量如下：

图 4-图 7 展示了相关的抽取结果和对生物多样性领域语义检索的支撑结果。

	identifier	nerName	nerIdentifier	nerType	literalIdentifier	literalZone	nerStart	nerEnd	language	annotatedTime	sourceDictionay
1	1c1f16cc6526b1af886d58ba9d18c1eb3-711	Illiciaceae	[100466]	family	1188180	Abstract_En	108	119	En	2014-05-06 18:01:35.887	PDBOntology
2	1c1f423b28d8bd4c94f49aa239a205eb3-711	Pinus	[119655]	genus	1207118	Abstract_En	1827	1832	En	2014-05-06 18:46:40.107	PDBOntology
3	1c1f7cd13077ed8b27a698cf9e64121b3-711	CAM	[100776]	photosynthesis	1200656	Abstract_En	720	723	En	2014-05-06 18:32:08.980	PDBOntology
4	1c1f7fe5f5562d217f5f46059367abbd3-711	fruit	[74459, 77005]	plant anatomical entity	97576	Abstract_En	847	852	En	2014-03-24 20:16:09.747	Ontology_AcceptedName
5	1c1f893d0da4aa38e58d648d124020e43-711	air	[28859]	environmental features and habitats	493815	Abstract_En	369	372	En	2014-03-24 20:09:41.420	Ontology_AcceptedName
6	1c1fb09e509787ad85d6e217f1a3cd0a3-711	tree	[100272]	habit	1189817	Abstract_En	533	537	En	2014-05-06 18:05:28.857	PDBOntology
7	1c1fc577eeb3fd38d1a0f4c8b8fe5c763-711	flower	[74215, 77067]	plant anatomical entity	392336	Abstract_En	807	813	En	2014-03-24 19:57:39.607	Ontology_AcceptedName
8	1c1fd00a8b102a9fac2be043727056553-711	polyacrylamide gel electrophoresis	[18137]	Laboratory Procedure	872257	Abstract_En	926	960	En	2014-03-24 20:11:34.107	Ontology_AcceptedName
9	1c1fd5643e82df384d66b1b86a5b8f193-711	gynodioecious	[120606]	plantFlowerSexuality	1199605	Abstract_En	396	409	En	2014-05-06 18:29:21.450	PDBOntology
10	1c1ff08ca5c5e40c48410422cfc2b2c03-711	water	[28245]	environmental features and habitats	488969	Abstract_En	74	79	En	2014-03-24 20:09:08.810	Ontology_AcceptedName
11	1c20c402af3a996396bd6767503977323-711	Calathea	[113630]	genus	1207930	Abstract_En	920	928	En	2014-05-06 18:48:41.590	PDBOntology
12	1c2293d5ebf1f23fa89eed9033ee1efb3-711	biota	[118360]	genus	1203263	Abstract_En	142	147	En	2014-05-06 18:39:28.230	PDBOntology
13	1c22dbcd77fafe67c4b5c94a169b92ab3-711	tree	[100272]	habit	1201832	Abstract_En	310	314	En	2014-05-06 18:35:11.340	PDBOntology
14	1c22f3289c92f2cd943e2af73f4cd7893-711	Arabidopsis	[58496]	genus	476296	Abstract_En	1522	1533	En	2014-03-24 20:04:33.247	Ontology_AcceptedName
15	1c22fc501be7317101ba6b275d07ac613-711	rice	species:ncbi:...	species:ncbi:4530	83442	Abstract_En	573	577	En	2014-03-24 19:49:10.793	NCBI
16	1c230443428af4deeaab25c19b32f78f3-711	Phoebe	[114415]	genus	1178927	Title_EN	35	41	En	2014-05-07 08:53:03.950	PDBOntology
17	1c23a70c79196895d1d5b631c4f1698b3-711	Quercus	[120001]	genus	1187812	Abstract_En	83	90	En	2014-05-06 18:00:43.760	PDBOntology
18	1c23b9dc27ae3ebdf31947b3e790409b3-711	Primula	[117847]	genus	1201740	Title_EN	45	52	En	2014-05-07 09:05:07.497	PDBOntology
19	1c23c0fb3344762f69d497125e260a7d3-711	free	[30686, 31557]	plantGynoeciumCarpelFusion|plan...	497987	Abstract_En	367	371	En	2014-03-24 20:10:21.277	Ontology_AcceptedName
20	1c23ce608354aa2f98c90c0c420f43153-711	Nicotiana	[61964]	genus	284748	Abstract_En	117	126	En	2014-03-24 19:55:03.920	Ontology_AcceptedName
21	1c23d9210e452e345a0209c2f5d31cb83-711	Arabidopsis thaliana	species:ncbi:...	species:ncbi:3702	100210	Abstract_En	0	20	En	2014-03-24 19:44:39.073	NCBI
22	1c243b3fd2d6b0695d0fe9f7ba21593c3-711	parenchyma	[78436]	plant anatomical entity	485844	Abstract_En	1160	1170	En	2014-03-24 20:07:27.420	Ontology_AcceptedName

图 4 生物多样性领域实体抽取结果示例

	identifier	nerName1	nerIdentifier1	nerName2	nerIdentifier2	relationType	relation	literalIdentifier	zone	annotatedTime
1	8bae15224b78...	French oceanographic research centre	3030235e66f3dead...	IFREMER	fa381f7a9f6614e8cc...	semantic_relation	apposition_relation_...	1196111	Abstract_En	2014-05-20 17:10:25.35
2	8bb25163d769...	SH	ec5704f0d56945d1...	Shimodaira-Hasegawa	342883ef1cb34ccc0...	semantic_relation	apposition_relation_...	1193127	Abstract_En	2014-05-20 17:10:33.18
3	8bb47cb7bdce...	subtropical woody dwarf bamboo	fbcc164889a6a06ff...	Nakai	02497673db92a38c...	semantic_relation	apposition_relation_...	1203822	Abstract_En	2014-05-20 17:07:30.96
4	8bb68f1704fad...	Sierra Madre Occidental	5c41d49097eea0d...	Oriental Socoriusco__Sierra_Madre	ee153c4e8e7e2e8c...	semantic_relation	apposition_relation_...	1178780	Abstract_En	2014-05-20 16:40:23.82
5	8bbb593de701...	P integrifolia L.	072f654cc737aa8c...	P. balbisii Lehm.	aa3949699bc62dc4...	semantic_relation	apposition_relation_...	1179007	Abstract_En	2014-05-20 17:12:27.01
6	8bc1fe28dca05...	Onobrychis	8fc9146e40ab39a1...	Alhagi Taverniera	df40b1e1a63707e6e...	semantic_relation	apposition_relation_...	1207684	Abstract_En	2014-05-20 17:13:01.46
7	8bc355a1da12...	information	bb3ccd5881d6514...	species	cd3d20b946198768...	semantic_relation	apposition_relation_...	1196134	Abstract_En	2014-05-20 17:01:29.97
8	8bc515af0398d...	ndhF	6b02b9d3db40885...	DNA sequence data	bbb42b6f90caf321c...	semantic_relation	apposition_relation_...	1207965	Abstract_En	2014-05-20 17:06:18.14
9	8bc5eace07a0...	ndhF	6b02b9d3db40885...	matK	4745ad295a52c4b3...	semantic_relation	apposition_relation_...	1200919	Abstract_En	2014-05-20 17:12:54.02
10	8bc7f2a202785...	leaf area	d4f8433741e3ba73...	leaves	79f7d2ef17285d43e...	semantic_relation	apposition_relation_...	1201783	Abstract_En	2014-05-20 16:37:08.26
11	8bc9fe23ee2ec...	Plantaginaceae	5f62b8dbb4ef4738...	Pseudolysimachion	a1a02052eb5cc69e...	semantic_relation	apposition_relation_...	1178313	Abstract_En	2014-05-20 17:07:27.66
12	8bcd9559504c...	seed bank germination	723eefd9e9f70d03...	seed dispersal	c6de66f094a23dd2ff...	semantic_relation	apposition_relation_...	1200247	Abstract_En	2014-05-20 16:48:32.28
13	8bd0544671ad...	Neogrolleoideae	ee2cccf858b71289...	relationships	eaa7c88f5c1246724...	semantic_relation	apposition_relation_...	1206711	Abstract_En	2014-05-20 17:13:41.76
14	8bd073b1f9aeb...	PS-1	e6ab54a04498e91...	Soybean_partial-female-sterile_mutant 1	6d8311de4515f5701...	semantic_relation	apposition_relation_...	1191122	Abstract_En	2014-05-20 17:13:05.92
15	8bd162c3b06c...	standing leaf_numbers	736f15f2f2edc4590...	significant_sun leaf_thickness	ecbcb3b5229b3c37...	semantic_relation	apposition_relation_...	1189919	Abstract_En	2014-05-20 16:38:20.41
16	8bd561bab288...	var	b2145aac704ce76...	varieties	1a1d4cb0fac4d95fe...	semantic_relation	apposition_relation_...	1205370	Abstract_En	2014-05-20 16:37:47.86
17	8bd8bcf9f183c...	ITS	fcdb76644228e946...	spacer	dddd37dfc736f0674...	semantic_relation	apposition_relation_...	1194070	Abstract_En	2014-05-20 17:13:41.89
18	8bdcdd3d7b14...	database management options	86e4019938c8cc2...	monitoring approaches	a8527b789cbcb72a...	semantic_relation	apposition_relation_...	1197641	Abstract_En	2014-05-20 17:13:39.48
19	8bdfc3d61c5ac...	Desmos saccopetaloides	aea2574f55fd4950f...	China	ae54a5c026f31ada0...	semantic_relation	apposition_relation_...	1187069	Abstract_En	2014-05-20 16:34:38.94
20	8be03debd9b5...	complete deletion	5df27a4d4ddb4900...	pAs1-FISH patterns	991fb99dbcc4f7087...	semantic_relation	apposition_relation_...	1206166	Abstract_En	2014-05-20 17:11:03.71

图 5 生物多样性领域语义关系抽取结果示例

从 30,665 篇文献中获得 133,922 条语法关系结果，类型均为 SPO 语法关系。

从 15,259 篇文献中获得 35,903 条语义关系结果，均为 apposition_relation 同位语关系。

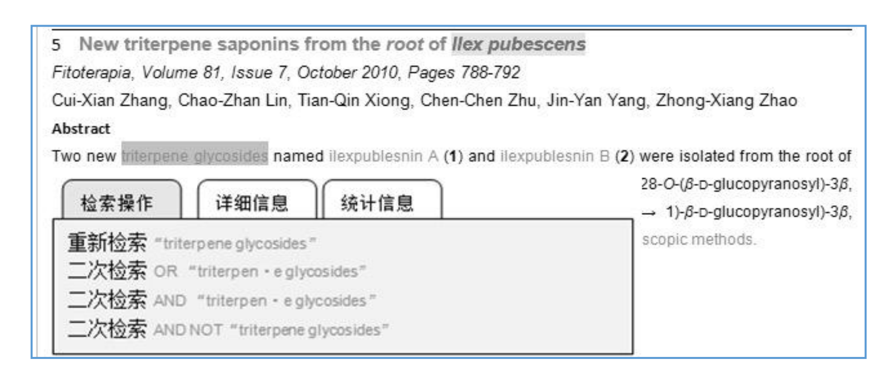


图 6 基于本体概念或实体的知识浏览、检索与统计分析功能

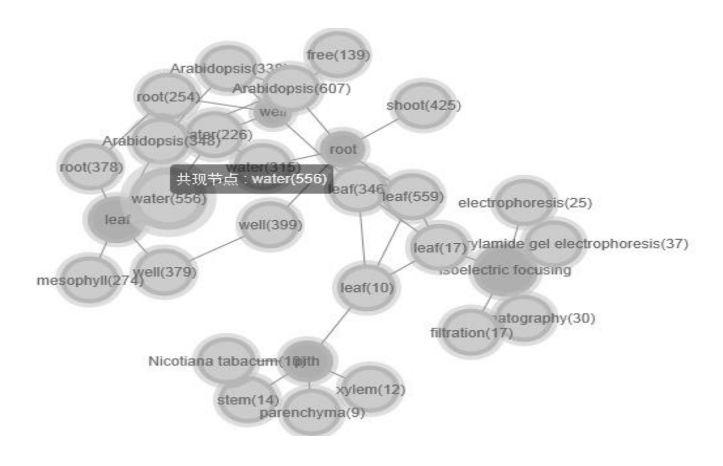


图 7 基于语义知识抽取的单篇文章共现关系知识图

4 结语

本文在对当前生物多样性信息抽取领域相关研究分析的基础上，结合中国科学院文献情报中心“建设生物多样性领域本体构建与语义组织应用示范平台”的实际要求，设计了生物多样性语义知识抽取框架，并利用十二五科技支撑计划“面向外文科技文献信息的知识组织体系建设与应用示范(STKOS) ”构建的植物多样性本体作为底层的词典，探索实现了相应的语义知识抽取方法，开发了相应的生物多样性示范平台。本研究更多从实际应用的层面探索了可工程化应用的知识组

织框架及知识识别的方法，因此，词典和人工撰写的规则是本研究中开展知识抽取的重要组成部分，正因为此，词典和人工规则本身所固有的局限性也在一定程度上限制了识别的完整性和准确性，在后续的研究中，针对各类型知识单元的精细化识别仍将是重要内容。

参考文献：

[1] Anne E T, Cui H, Mozzherin D. Applications of Natural Language Processing in Biodiversity Science[J]. Hindawi Publishing Corporation Advances in Bioinformatics,2012, doi:10.1155/2012/391574

[2] Naderi N., Kappler T, Baker J.O C, et al. OrganismTagger: Detection, Normalization and Grounding of Organism Entities in Biomedical Documents[J].Bioinformatics. 2011,27(19):2721-9

[3] Species[EB/OL].[2016-04-12].http://en.wikipedia.org/wiki/Species

[4] Gerner M, Nenadic G,Bergman C M.LINNAEUS: A Species Name Identification System for Biomedical Literature[J],BMC Bioinformatics, 2010(11):85

[5]The NCBI Taxonomy Homepage[EB/OL].[2016-04-12]. http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/

[6]Page RDM. (2013) BioNames: linking taxonomy, texts, and trees. PeerJ 1:e190 http://dx.doi.org/10.7717/peerj.190.

[7]Species 2000[EB/OL].[2016-04-12]. http://www.catalogueoflife.org/annual-checklist/2014/

[8] Akella L M, Norton C N, Miller H. NetiNeti: discovery of scientific names from text using machine learning methods[J]. BMC Bioinformatics. 2012(13):211

[9] The OrganismTagger System[EB/OL].[2016-04-12]. http://www.semanticsoftware.info/organism-tagger

[10]Koning D, Sarlar I N, Moritz T.Taxongrab: Extracting Taxonomic Names from Text[J].Biodiversity Informatics, 2005(2):79-82

[11] Taylor A. Extracting Knowledge from Biological Descriptions [C]. In: Proceedings of the 2nd International Conference on Building and Sharing Very Large-Scale Knowledge Bases. 1995: 114-119

[12]Tang X, Heidorn P B. Using Automatically Extracted Information in Species Page Retrieval [OL]. [2016-03-29]. http://www.tdwg.org/proceedings/article/view/195/

[13] Cui H. CharaParser for Fine-grained Semantic Annotation of Organism Morphological Descriptions [J]. Journal of the American Society for Information Science and Technology, 2012, 63(4): 738-754

[14]段宇锋, 黄思思. 中文植物物种多样性描述文本的信息抽取研究[J]. 现代图书情报技术, 2016, 32(1): 87-96. Duan Y F, Huang S S. Information Extraction from Chinese Plant Species Diversity Description Text[J]. New Technology of Library and Information Service, 2016, 32(1): 87-96.

[15] Li C, Liakata M, Rebholz-Schuhmann D. Biological network extraction from scientific literature: state of the art and challenges[J]. Briefings in bioinformatics. 2013(2):6

[16] Skusa A, Rüegg A, Köhler J. Extraction of biological interaction networks from scientific literature[J]. Briefings in Bioinformatics 2005,6(3): 263-276.

[17] 许哲平, 崔金钟, 覃海宁, 等. 中国生物多样性 e-Science 平台建设构想[J]. 生物多样性, 2010, 18(5): 480-488. Xu Z P, Cui J Z, Qin H N, Ma K P. On the architecture of biodiversity e-Science infrastructure in China[J]. Biodiversity Science, 2010, 18(5): 480-488.

(通讯作者: 刘建华，ORCID：0000-0002-4003-8834，E-mail:liujh@mail.las.ac.cn)

作者贡献说明:

刘建华：提出论文整体框架，完成语义知识抽取框架的设计，参与实现知识抽取的实现开发，完成论文主体内容的撰写，对最终版本部分内容进行校对、修改完善。
王颖：参与设计语义知识抽取框架和知识抽取开发的语料准备、存储结构设计。
张智雄：对研究过程进行指导，对论文内容提出修改意见。
李传席：主要负责知识抽取功能的实现开发，并提供开发文档。

利益冲突声明：

所有作者声明不存在利益冲突关系。

支撑数据：

[1]